به نام خدا

دانشگاه تهران

دانشکده مهندسی برق و کامپیوتر

مقدمه‌ای بر استنباط آماری

گزارش تمرین اول

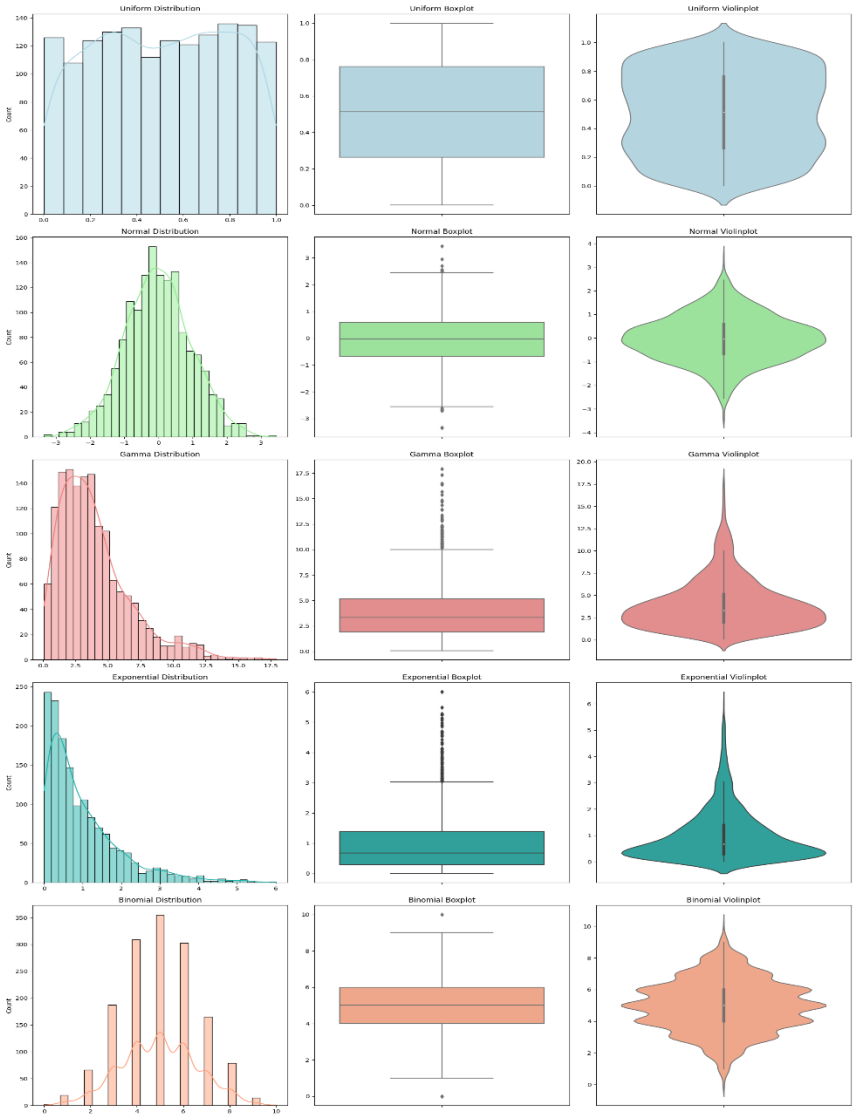
نام و نام خانوادگی	سید مهدی موسوی
شماره دانشجویی	۸۱۰۱۰۲۲۶۴
تاریخ ارسال گزارش	۱۴۰۲/۷/۱۸

فهرست گزارش سوالات

سوال دهم:

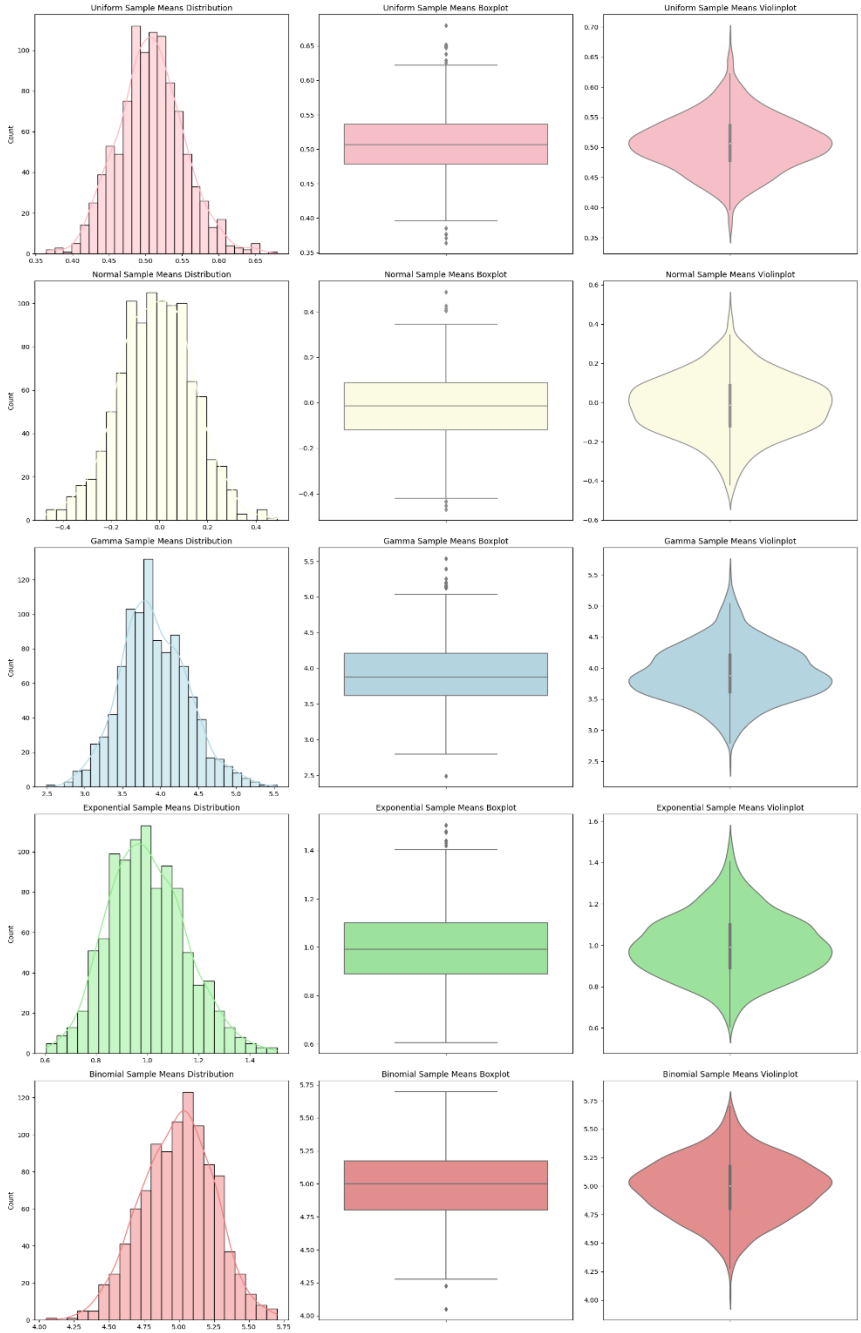
تولید اعداد تصادفی

در اولین قسمت از این سوال به تولید اعداد تصادفی از توزیع های متفاوت می پردازیم تا شکل آنها را ببینیم. توزیع های مد نظر هم یکنواخت در بازه ۰تا ۱, نرمال با میانگین ۰ و واریانس ۱, گاما با پارامتر ۲ , نمایی با متغیر ۱ و در نهایت چند جمله ای با احتمال ۰.۵ است. شکل توزیع های گفته شده در شکل ۱-۱۰ آورده شده است. شکل های زیر به کمک ۱۵۰۰ نقطه به دست آمده است و همانطور که دیده میشود از bar plot, box plot و violin plot استفاده شده است.



نمودار ۱-۱۰ : شکل توزیع های تولید شده

در ادامه ۱۰۰۰ دسته ۴۰ تایی از توزیع های نامبرده شده تولید میکنیم و از آنها میانگین میگیریم و توزیع این ۱۰۰۰ میانگین را به تصویر می کشیم. طبق قانون اعداد بزرگ باید این توزیع به شکل نرمال باشد و همانطور که در نمودار ۲-۱۰ دیده میشود این توزیع ها به شکل نرمال در آمده اند. در مورد توزیع های مثل نمایی چون توزیع چولگی دارد توزیع میانگین ها کمی از شکل نرمال فاصله دارد اما در نهایت باز هم میتوان با تقریب خوبی آن را نمایی در نظر گرفت.

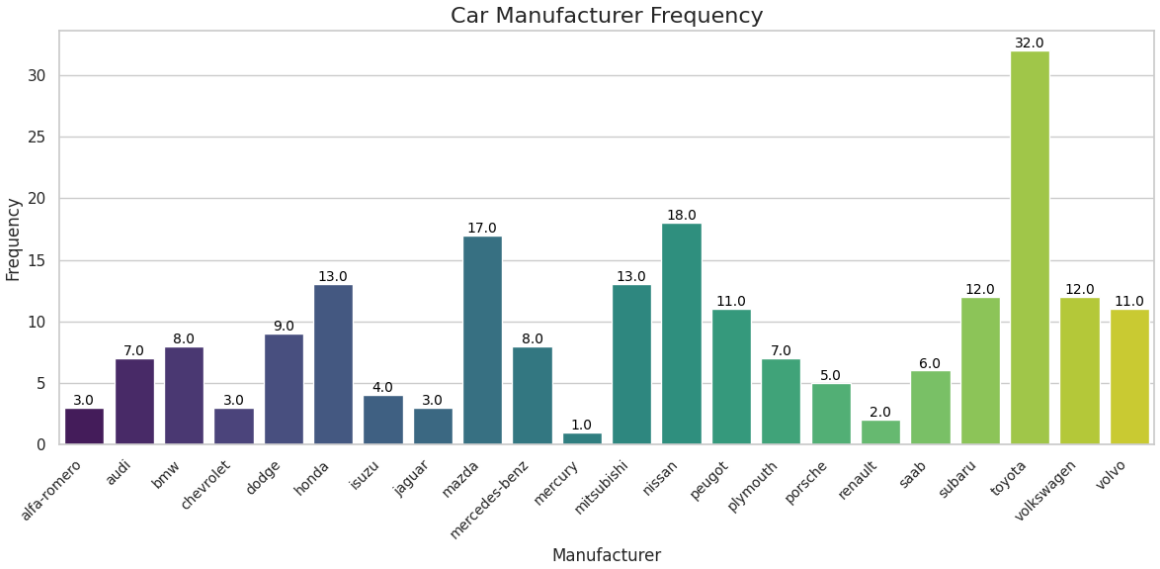


شکل ۲-۱۰ : نمودار های مختلف برای میانگین دسته های ۴۰ تایی از توزیع های مختلف

کار با دادگان ماشین

در قسمت بعد از سوال به کار با دادگان داده شده میپردازیم. در نگاه اول به این داده ها مشاهده کردیم که در بعضی از ردیف ها به دلیل نبود اطلاعات ? گزاشته شده است پس در اولین قدم این کاراکتر را به مقدار متناسب یعنی nan تغییر میدهیم. حال به بررسی اطلاعات ستون های مختلف میپردازیم که دیده میشود به غیر از ستون normalized-losses باقی ستون ها دارای تعداد کمی مقدار نامعین هستند پس تمام ردیف هایی که در ستونی به غیر از ستون normalized-losses مقدار نامشخص دارند را از داده حذف میکنیم. در این مرحله تعداد ردیف ها از ۲۰۵ به ۱۹۳ کاهش می یابد. همچنین باید نوع همه ستون ها را به غیر از ستونی که دارای مقادیر نامشخص است به مقدار مناسب تغییر بدهیم که هر ستون به غیر از ستون نامبرده شده به یکی از دسته های categorical,int و float تبدیل شد.

در مرحله بعد به بررسی سازندگان خودرو های در دست می پردازیم. با توجه به داده ها تویوتا بیشترین تکرار را در بین داده های در دست دارد.همچنین برای توزیع تعداد بر حسب سازنده خودروها به نمودار ۳-۱۰ خواهیم رسید.

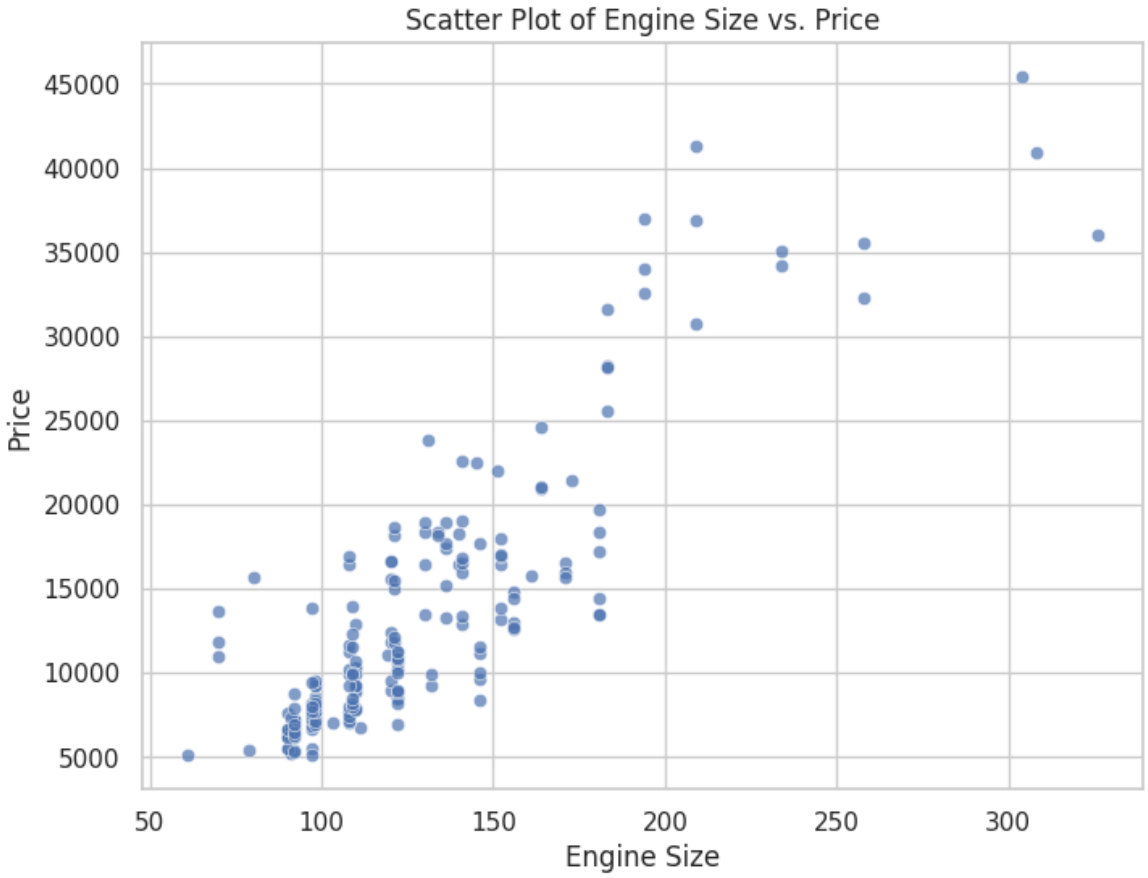


نمودار ۳-۱۰ : نمودار تعداد برای هر سازنده ماشین

برای محاسبه معیار های پراکندگی همانطور که خواسته شده است از سه معیار استفاده کردیم. با توجه به معیار skewness میتوان گفت که سه معیار نسبت فشردگی, قیمت و همچنین اندازه موتور بیشترین چولگی را دارند که همگی چولگی راست دارند. همچنین میتوان دید که معیار stroke تنها معیاری است که مقدار چولگی چپ قابل توجهی دارد. همچنین با بررسی معیار دیگر یعنی پخی میتوان دید که دقیقا سه معیار ذکر شده با مقدار چولگی راست زیاد بسیار متمرکز حول مرکزشان هستند که میتوان از آن

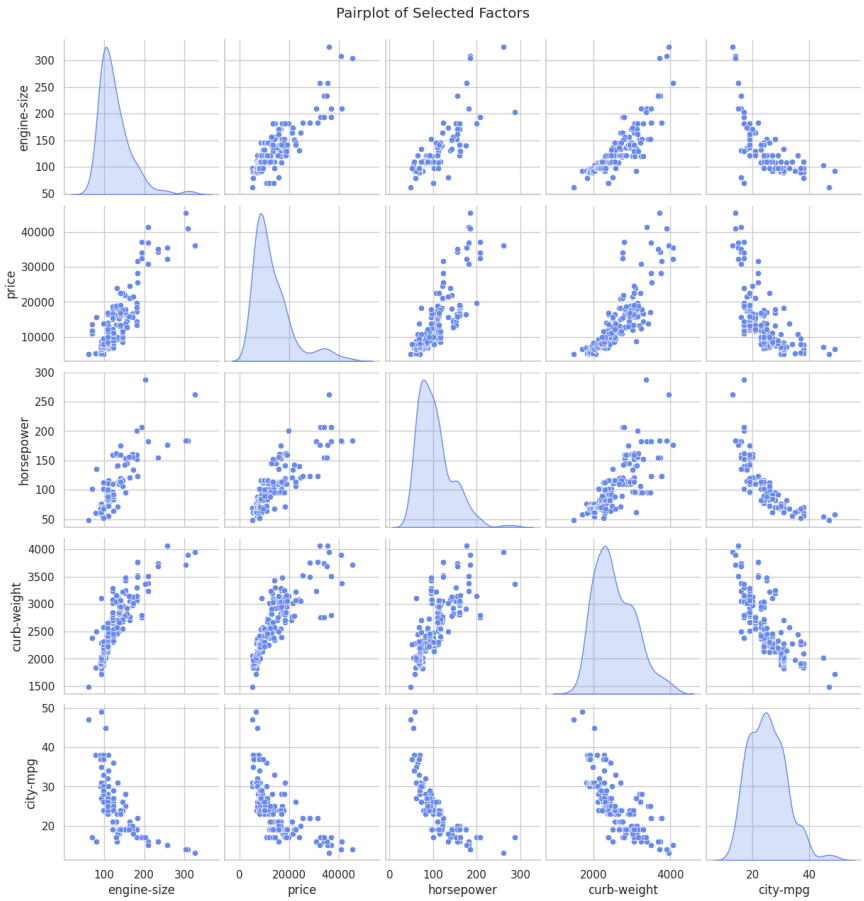
نتیجه گرفت که این سه معیار دارای مقدار هایی با اعداد خیلی بزرگ و تعداد کم دارد . همچنین متغیر bore پخ ترین متغیر در دادگان در دست است.

در ادامه به بررسی رابطه اندازه موتور و قیمت ماشین می پردازیم. همانطور که در نمودار ۴-۱۰ دیده میشود با افزایش اندازه موتور میتوان دید که به احتمال زیاد قیمت ماشین نیز افزایش می یابد. به عبارت دیگر انتظار ضریب همبستگی عددی مثبت و نزدیک یک باشد.



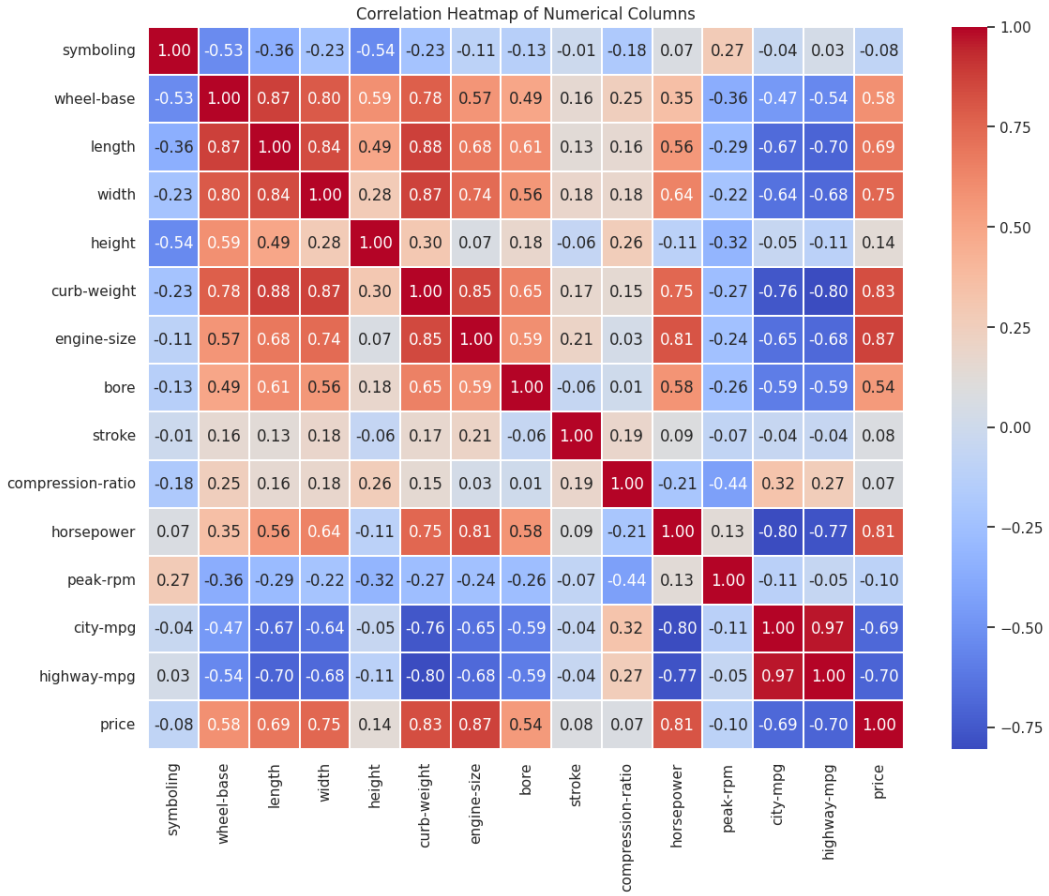
نمودار ۴-۱۰ : نمودار توزیع قیمت بر حسب اندازه موتور

حال به رسم pairplot برای متغیر های قیمت , مصرف سوخت , قدرت موتور ,اندازه موتور و همچنین وزن میپردازیم که در نمودار ۵-۱۰ دیده میشود. در این نمودار ۵ متغیر گفته شده را دو به دو با هم بررسی کرده ایم و در قطر آن توزیع خود متغیر دیده می شود. میتوان به عنوان مثال دید که اندازه موتور با مقدار توان تولیدی موتور بر حسب اسب بخار همبستگی مثبت احتمالا دارد .



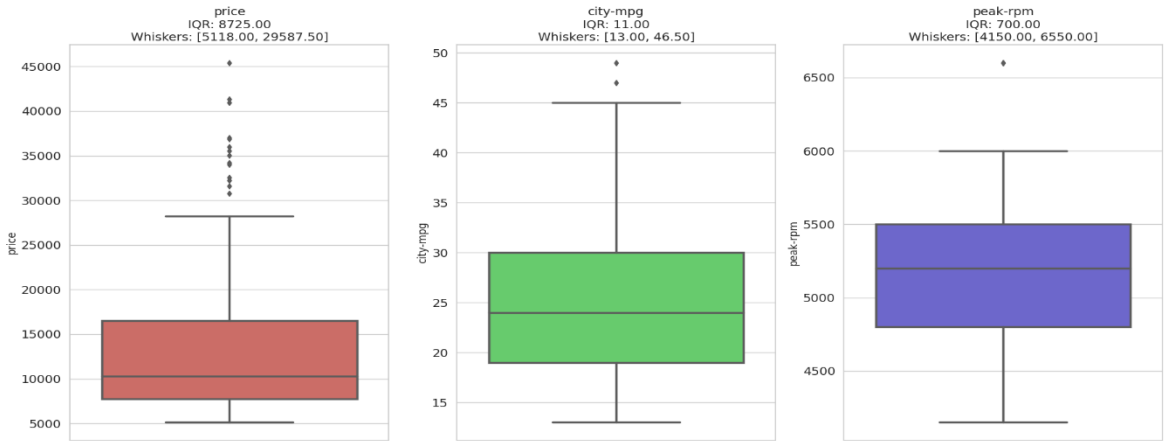
نمودار ۵-۱۰ : نمودار pairplot برای متغیر های قیمت , مصرف سوخت , قدرت موتور ,اندازه موتور و وزن

برای نمایش همبستگی میان متغیر ها هم از هیت مپ بین متغیر های مد نظر یعنی متغیر های عددی استفاده میکنیم. همانطور که از قبل انتظار داشتیم و در این هیت مپ هم دیده میشود دو متغیر اندازه موتور و قدرت آن به هم وابستگی زیادی دارند و ضریب همبستگی بین این دو متغیر ۰.۸۱ است که نشان دهنده این است که این دو متغیر به طور توامان زیاد میشوند.



نمودار ۶-۱۰ : هیت مپ ضریب همبستگی متغیر های عددی

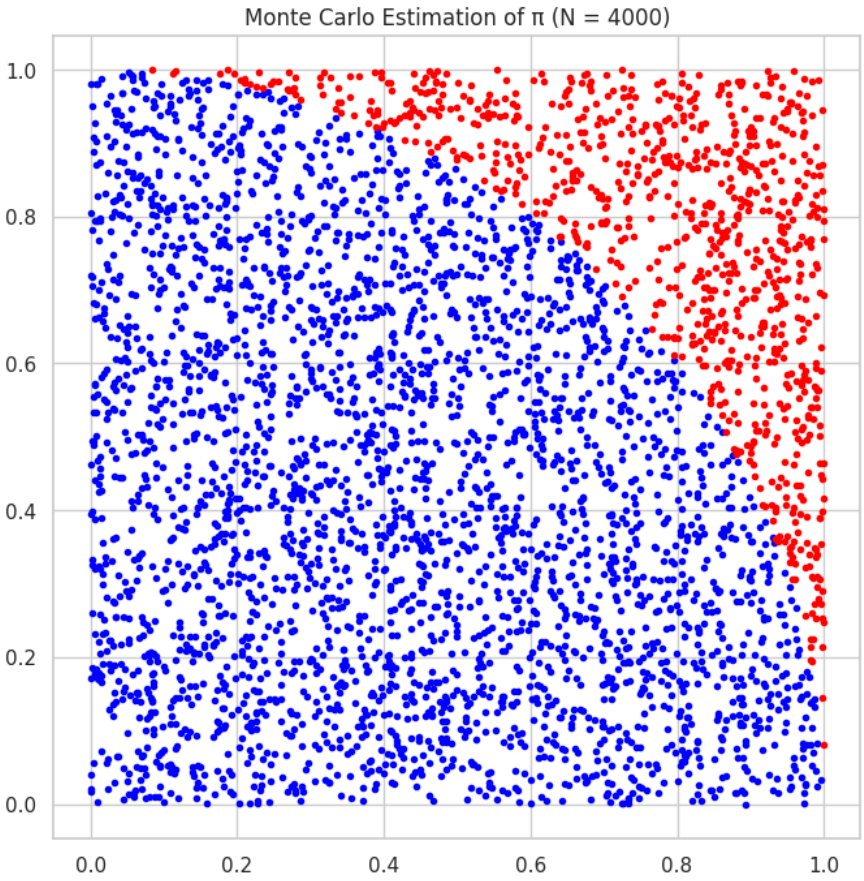
در آخرین مرحله از بررسی داده های دادگان در دست هم به رسم نمودار box plot برای سه متغیر قیمت , مصرف سوخت و حداکثر دور موتور میپردازیم. آماره های خواسته شده در عنوان نمودار ها وجود دارد



نمودار ۷-۱۰ : نمودار box plot برای سه متغیر قیمت , مصرف سوخت و حداکثر دور موتور

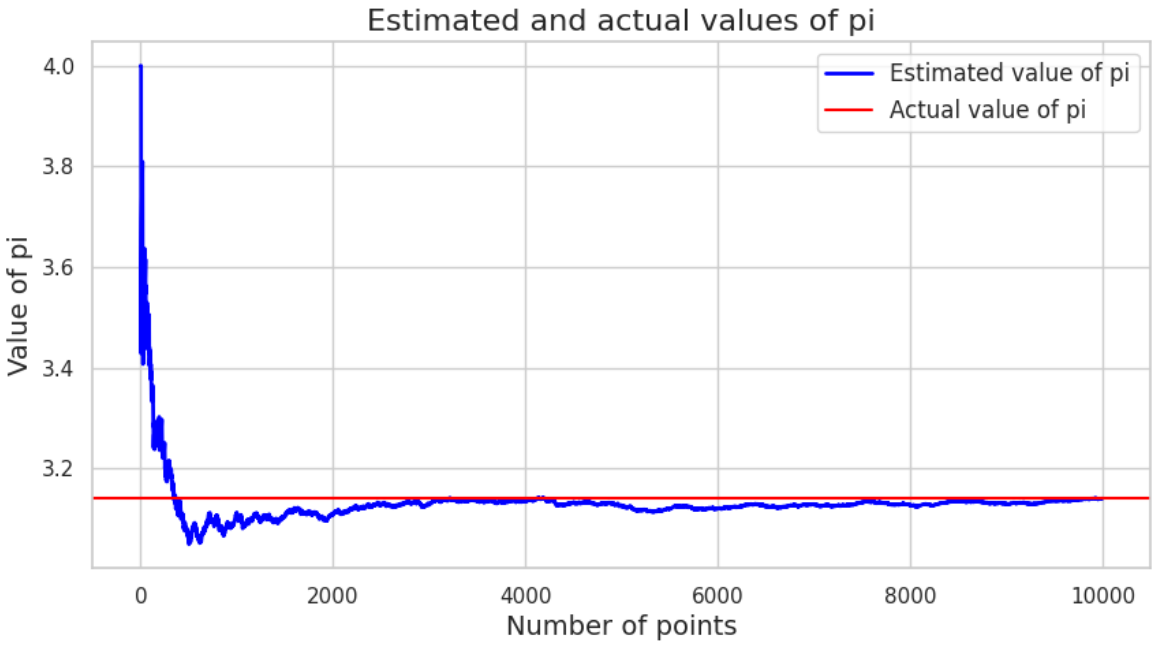
سوال دوازدهم:

در این سوال به تخمین مقدار پی با استفاده روش مونته کارلو میپردازیم. اساس این روش این است که تعداد زیادی نقطه تصادفی تولید کنیم و ببینیم چند درصد ازین نقاط درون ربع دایره قرار گرفته است.در نمودار ۱-۱۲ دیده میشود که به ازای ۴۰۰۰ نقطه تولید شده الگوریتم به چه صورتی عمل کرده است .



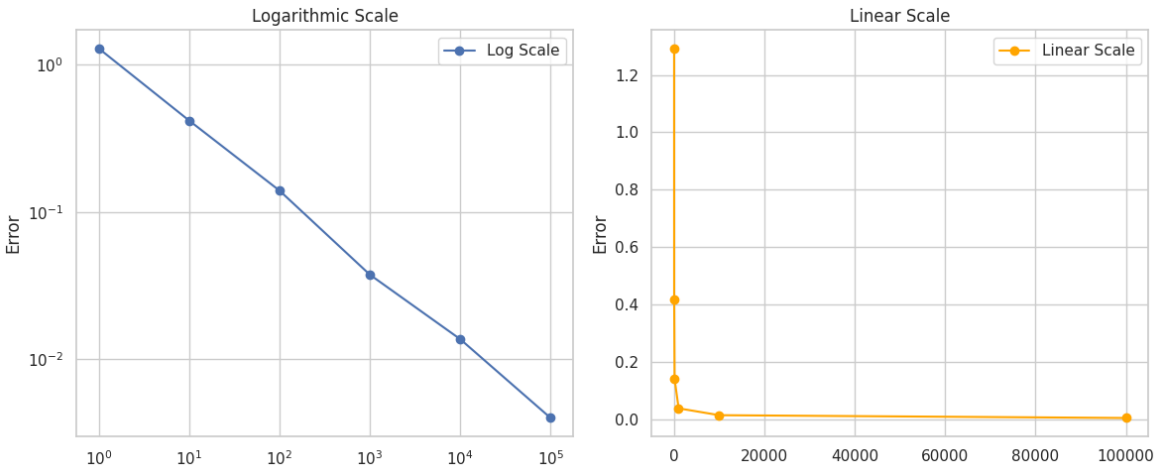
نمودار ۱-۱۲ : توزیع نقاط به ازای شبیه سازی برای ۴۰۰۰ نقطه

حال به بررسی مقدار خطای این روش برای تعداد مختف نقطه می پردازیم. همانطور که در نمودار ۲-۱۲ دیده میشود به ازای مقادیر کوچک برای تعداد نقطه خطای زیادی داشتیم و این مقدار به طور تدریجی کم شده تا اینکه برای مقدار نهایی تعداد نقاط یعنی ۱۰۰۰۰ مقدار خطا به حدود ۰.۰۰۱ رسیده است.



نمودار ۲-۱۲: مقدار واقعی پی و مقدار تخمینی در طول آزمایش

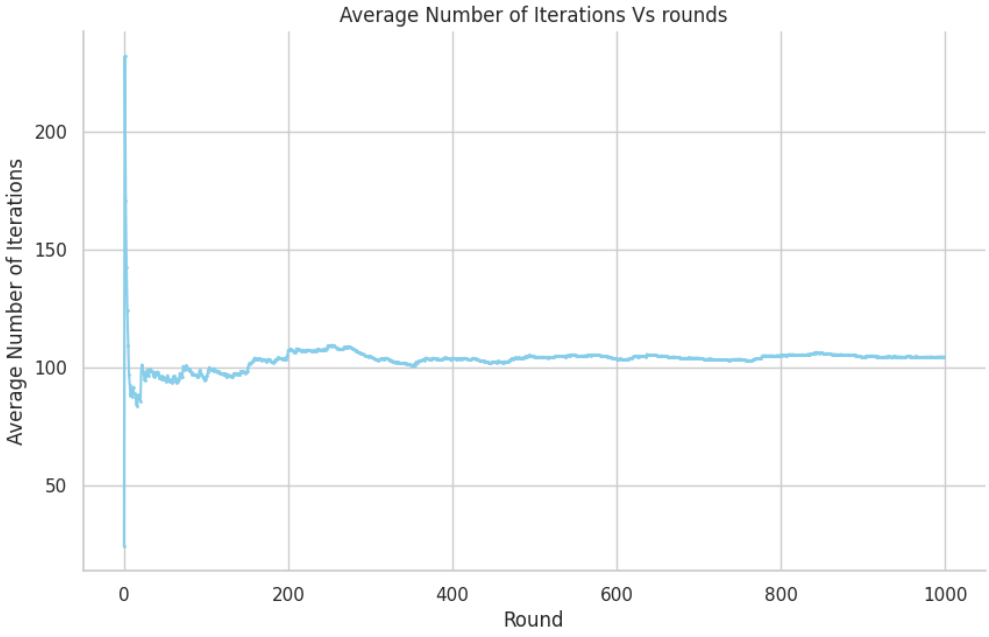
همچنین در ادامه الگوریتم را برای تعداد نقاط مختلف (توان های ۰ تا ۵ برای عدد ۱۰) هر کدام ۱۰۰ بار تکرار کردیم تا به تخمین مناسبی برای خطا در هر کدام از این نقاط برسیم. نتیجه در نمودار ۳-۱۲ قابل مشاهده است.



نمودار ۳-۱۲: مقدار خطا به ازای تعداد نقاط در دو نمودار لگاریتمی و خطی

سوال سیزدهم:

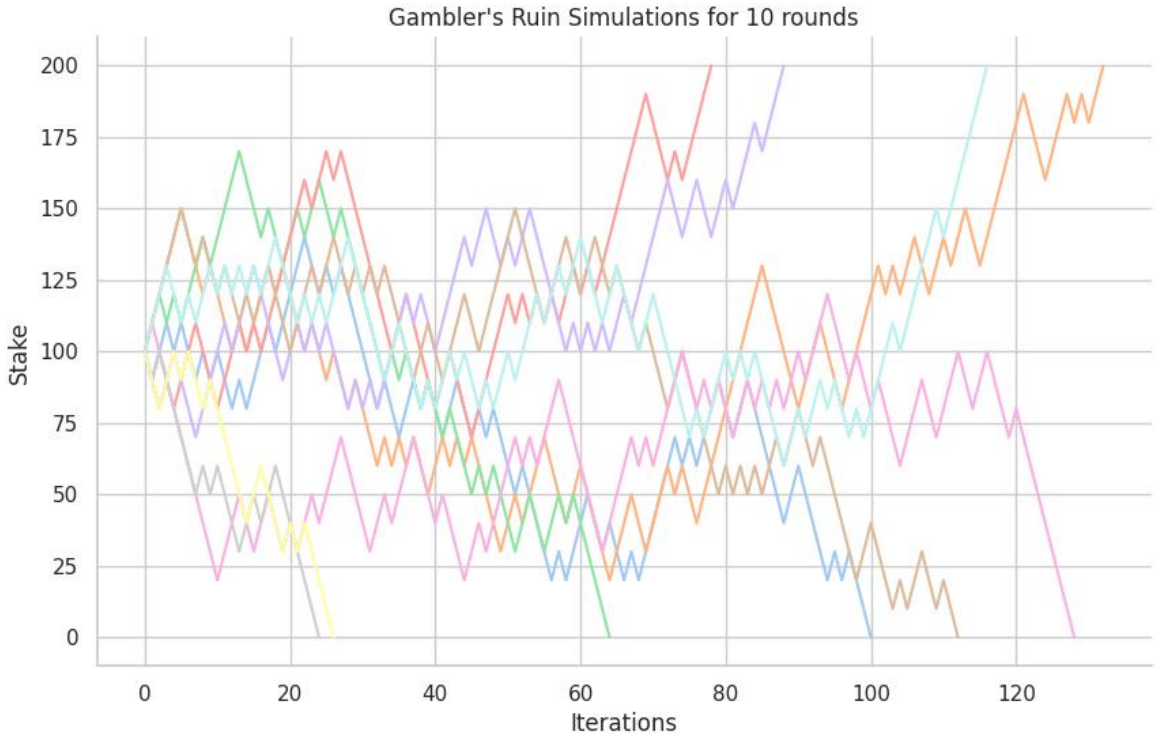
در این سوال به شبیه سازی یک قمار باز میپردازیم . با اجرای برنامه به ازای متغیر های متفاوت دیده شد که تغییر مقدار احتمال برد از ۰.۵ به مقداری کمی بیشتر باعث میشود که احتمال موفقیت بسیار بیشتر شود و بالعکس یعنی احتمال برد پایینتر از ۰.۵ باعث میشود احتمال موفقیت کلی بسیار پایین بیاید. همچنین نسبت فاصله مقدار اولیه با مقدار مد نظر به فاصله مقدار اولیه تا ۰ یکی دیگر از عوامل دخیل است به این صورت که اگر مقدار مد نظر تا مقدار اولیه فاصله کمتری داشته باشد به نسبت مقدار اولیه تا ۰ احتمال موفقیت بیشتر میشود. همچنین هر چقدر کسر بزرگتری از کل پول در قمار شرکت داده شود نویز رفتاری بیشتر میشود. آزمایش انجام شده توسط من با مقدار اولیه ۱۰۰ , مقدار نهایی ۲۰۰, احتمال برد ۰.۵ و مقدار شرط در هر دور ۱۰ بود. در نهایت همانطورکه در نمودار ۱-۱۳ دیده میشود بعد از کمی ناپایداری میتوان گفت که به طور میانگین با ۱۰۲ iteration به جواب مد نظر همگرا شده ایم و برای مقادیر اولیه مطرح شده احتمال پیروزی ۴۹ درصد به دست آمد.



نمودار ۱-۱۳ : میانگین تعداد iteration برحسب round

همینطور در نمودار ۲-۱۳ مقدار پول در دست ده قمارباز در طول این شبیه سازی مشاهده دیده میشود. همانطور که دیده میشود هیچ کدام از این ده قمار باز از ۱۲۰ بار بیشتر قمار نکرده اند و همگی

قبل از این تعداد قمار یا ورشکست شده اند یا با پیروزی از بازی خارج شده اند.



نمودار ۲-۱۳ : نمودار تغییرات پول در دست ده قمار باز در حین شبیه سازی